Buoy Project Report

Fall 2020 MA615 Assignment 1

Yuelin Jiang

9/25/2020

Buoy data and Global Warming

This report is a limited data exploratory analysis of Buoy data from the NOAA National Data Buoy Center, NDBC Station 44013. The buoy data is collected every hour from December 1987 to December 2016. My task is to extrapolate useful variables and try to measure global warming based on these 30 year data.

My Method

Abstract data of year 1987 to 2016 from the NDBC website, only keep the columns of year, month, day, hours, air temperature and water temperature. Transform temperatures to doubles and year month day hours to integers.

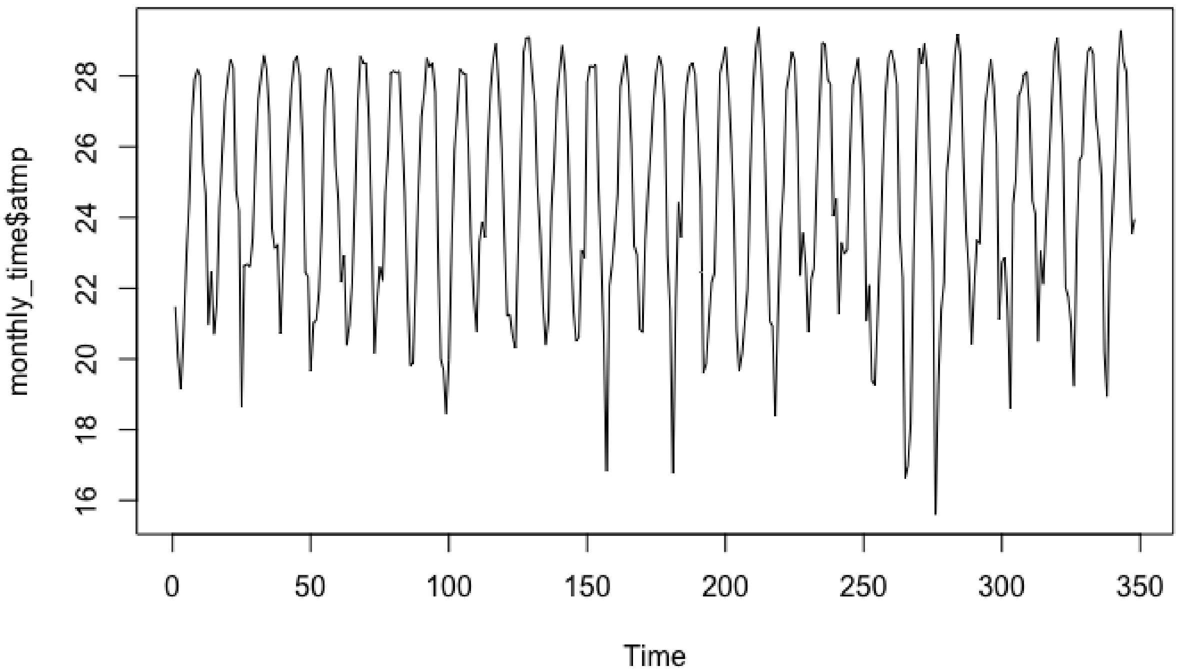
I only explore air temperature and water temperature at noon(hh=12), as the representative of the highest temperature for the day. Then I calculate the average monthly temperature for each year, using these daily noon temperatures. The reason why I only select temperature at noon is to avoid daily fluctuations which will affect the mean, so I pick one hour of the day to be consistent.

Analysis was conducted in R (R Core Team, 2014), data cleaning was performed using package lubridate (Grolemund, Wickham 2011) and package tidyverse (Wickham et al., 2019)). Time series analysis was conducted using package astsa (Stoffer, 2020), linear regressions were conducted using package rstanarm (Goodrich B, Gabry J, Ali I & Brilleman S., 2020) and figures were produced using the package ggplot2 (Wickham, 2009). This report was compiled by Knitr (Yihui Xie, 2020).

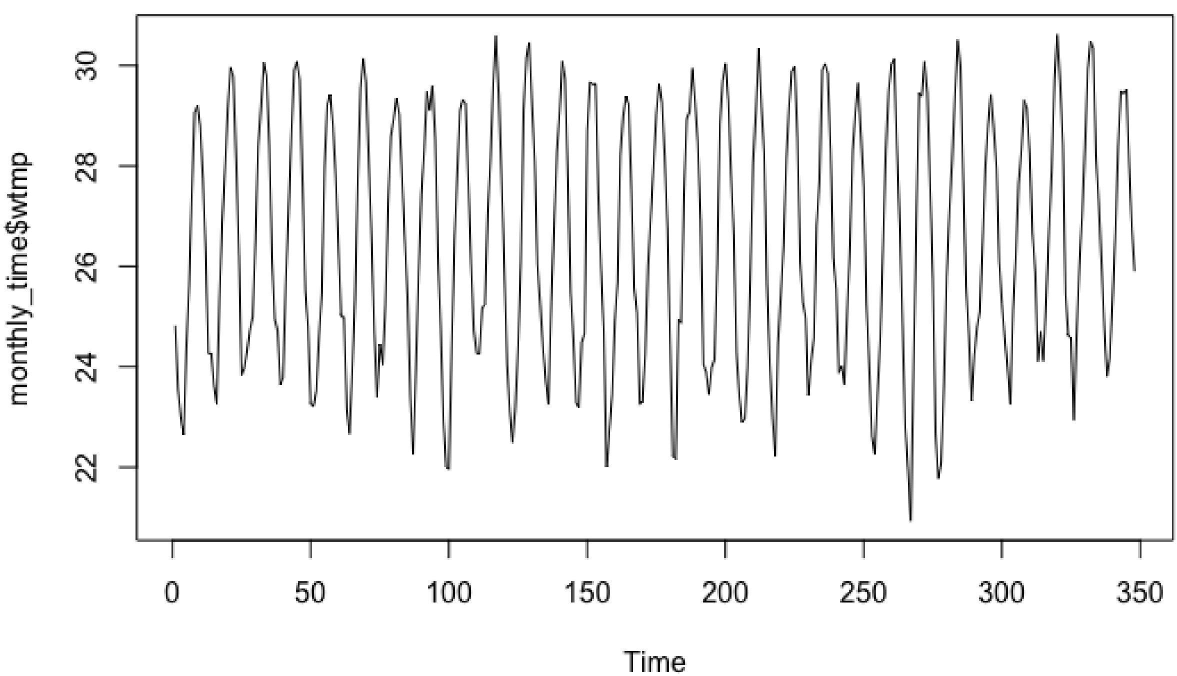
Data Analysis

Time Series Model

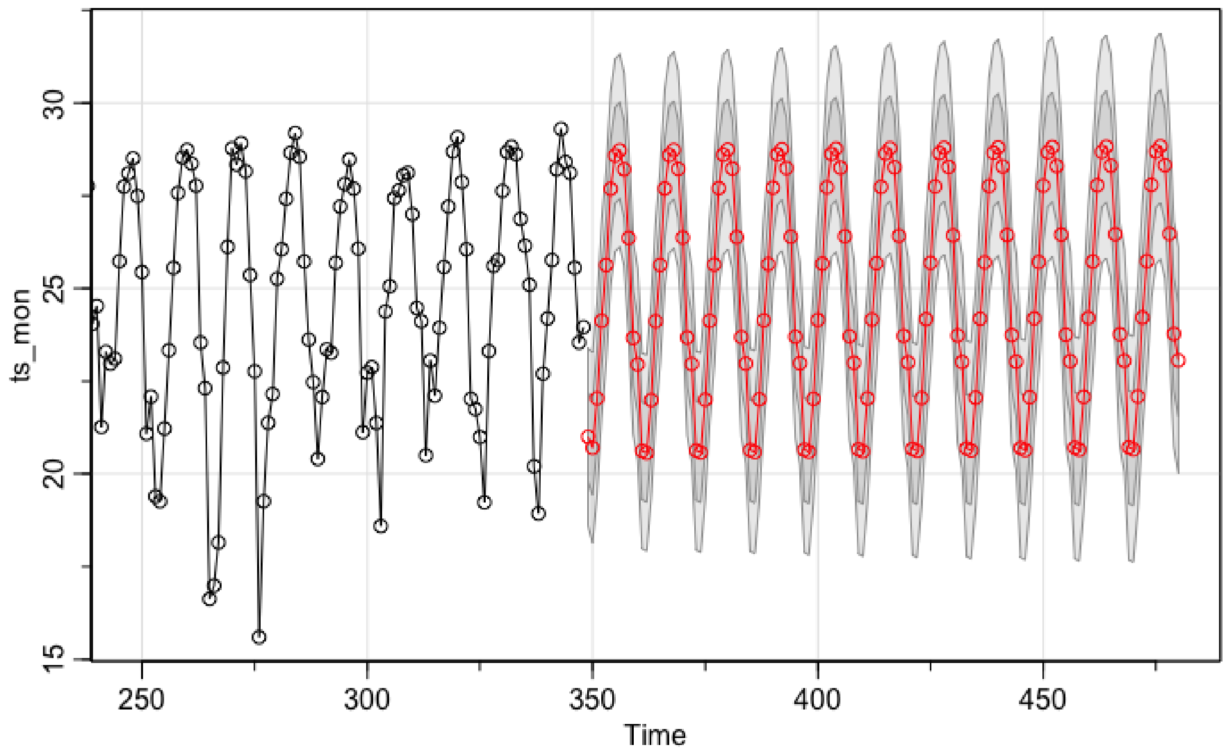
First I plot monthly average air temperatures over the 30 year period(1987-2016), shown below:



Then I do the same for monthly water temperatures over 30 year period(1987-2016), shown below:



We can observe that the overall trend of temperatures is hard to interpret because of the seasonal changes. Using the time series sarima forecast function to predict the seasonal trend we have shown below.



The forecasting result is not clear whether the overall temperatures are rising or decreasing, and there are a lot of fluctuations too (gray area), so it's not prudent to make any judgment based on this model alone.

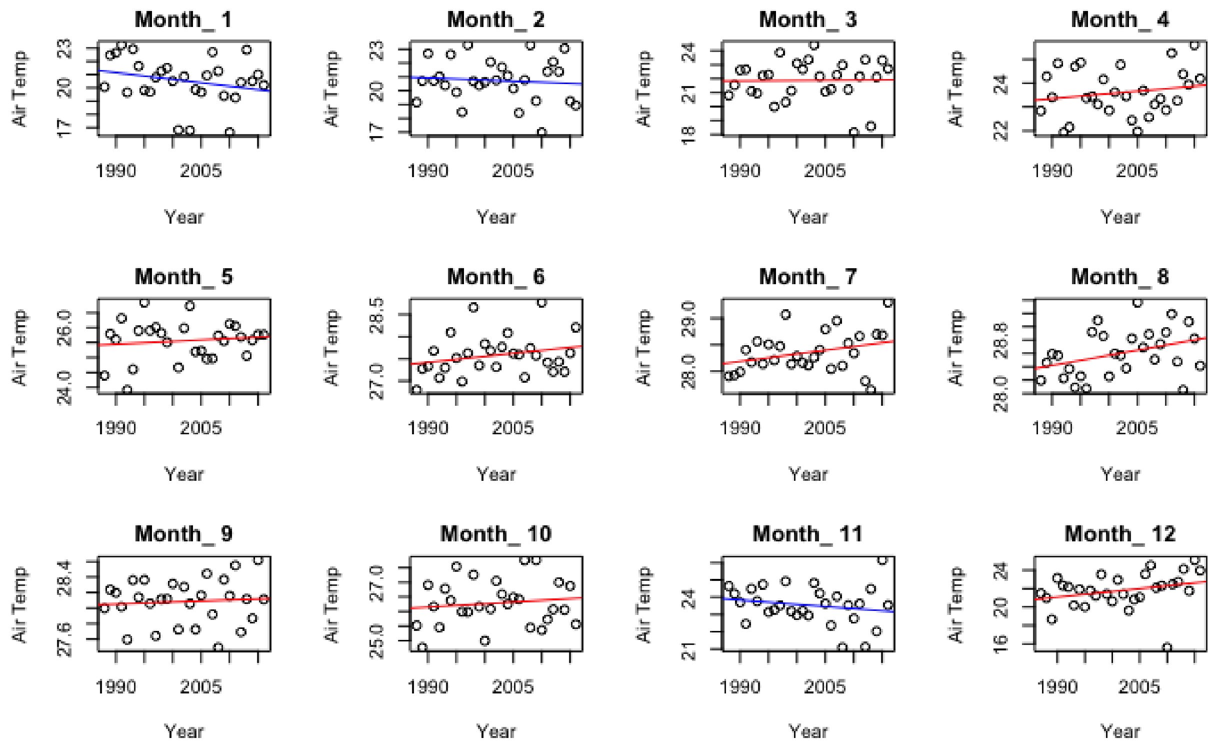
Linear Model

To overcome this seasonal change within each year, I proceed to run separate regressions for each month, so I have 12 regressions for each month, where temperature is a function of year.

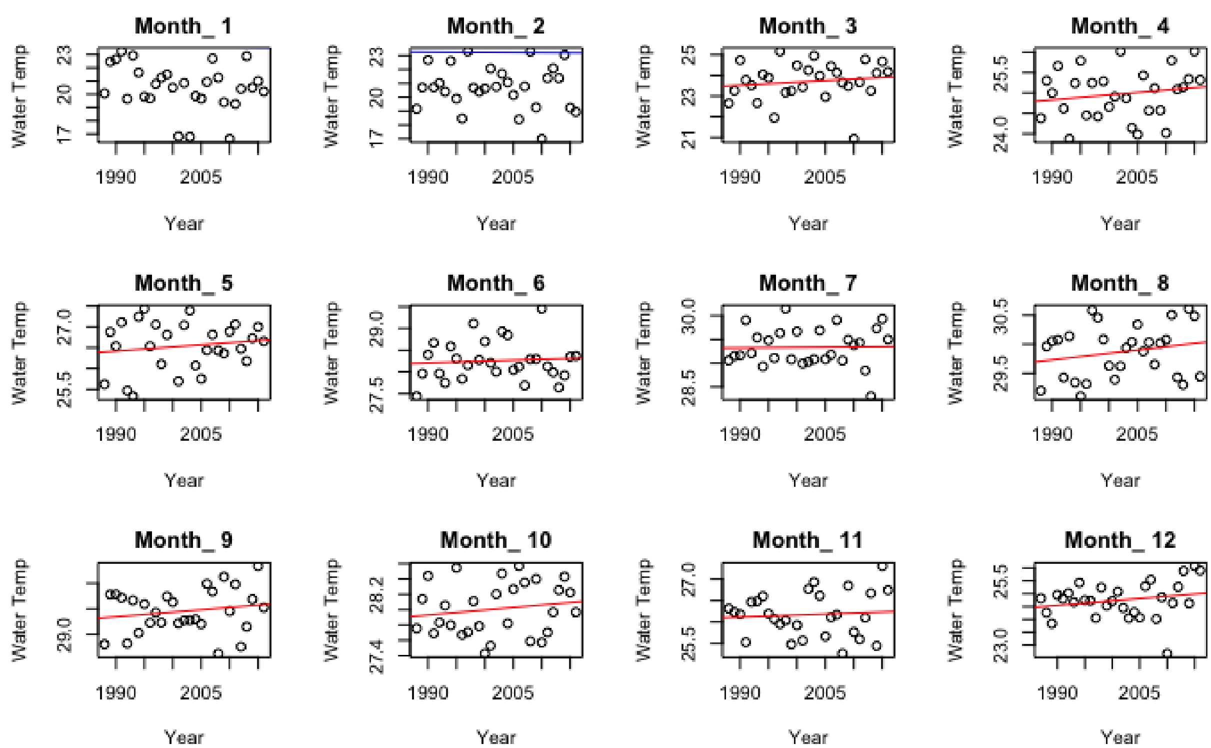
If there is indeed global warming, I should expect most of the months' average temperatures rising over the years. If there is climate change where weathers become more extreme, then I should observe summer temperatures rising and winter temperatures decreasing.

After cleaning and preparing the data, I used a for loop to do each linear regression and plotting. For example, at each loop for i, I group all the months that equals to i into a separate dataframe. Then I used lm() to regress temperature against year.

Depending on the coefficient of the slope, I draw the abline() red if the coefficient >0(increasing temp), and blue if the coefficient <0 (decreasing temp). The results for Each Month's Air Temperature Trend for the Year 1987-2016 are shown below:



Repeat the above process, and the results for Each Month's Water Temperature Trend for the Year 1987-2016 are shown below:



Conclusion

Based on the time series models, the seasonal trend of air and water temperatures is not clear.

Based on the linear regression models, for both air and water temperatures, most months have rising temperatures over the last 30 years. January, February and November have decreasing temperatures. This is within expectations, since colder winters are consistent with Climate change.

Although we cannot 100% confirm global warming in the past 30 years, we can conclude that for the past 30 years, around Boston area where this buoy collects data, Summers, Springs and Autumns in Boston are getting hotter, while Winters are getting colder.

This analysis certainly contains its limitations. The data is collected at one location on this planet, and my analysis only accounts noon temperatures. Further investigations could include more data collections around the globe, include more variables like CO2 level, and study daily, seasonal, and yearly temperature fluctuations.

Reference

Acknowledgements

Thanks to my teammates Zijie Huang, Ziyi Bai, Xiaozhou Lu, and Yuxin Zeng who helped me to improve my method and coding.

Package Reference

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Goodrich B, Gabry J, Ali I & Brilleman S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1 https://mc-stan.org/rstanarm.

Brilleman SL, Crowther MJ, Moreno-Betancur M, Buros Novik J & Wolfe R. Joint longitudinal and time-to-event models via Stan. StanCon 2018. 10-12 Jan 2018. Pacific Grove, CA, USA. https://github.com/stan-dev/stancon_talks/

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

David Stoffer (2020). astsa: Applied Statistical Time Series Analysis. R package version 1.10. https://CRAN.R-project.org/package=astsa

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595